AFRL-SR-AR-TR-09-0088

REPORT DOCUMENTATION PAGE

1. REPORT DATE *(DD-MM-YYYY)*
28-02-2009

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
March 2006 to September 2008

4. TITLE AND SUBTITLE
PROSODIC STRESS, INFORMATION, AND INTELLIGIBILITY OF SPEECH IN NOISE

5a. CONTRACT NUMBER
FA9550-06-1-0137

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)
Pierre L. Divenyi, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
East Bay Institute for Research and Education
PO Box 2339
Martinez, CA 94553

8. PERFORMING ORGANIZATION REPORT NUMBER
607-AFOSR-FR

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
875 M Randolph St
Arlington, VA 22203
Dr Willard Larkin

10. SPONSOR/MONITOR'S ACRONYM(S)
EBIRE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release

20090325289

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Prosodic stress increases the salience of stressed syllables. The project investigated whether this property of speech is used by listeners for the understanding of spoken sentences presented in noise. Stressed syllables in the 720-sentence IEEE corpus were marked and envelope contours were generated to increase or decrease the level of speech-spectrum noise in synchrony with the occurrence of stressed syllables. Data from ten normal-hearing young listeners indicate that signal-to-noise ratio (SNR) of stressed syllables alone is a good predictor of the intelligibility of the whole sentence. A computational model was also developed by decomposing speech into eight low-rate basis functions inspired by articulatory gestures. This model was applied to test the hypothesis that the slowly varying basis functions may reflect the way listeners recover low-SNR or otherwise distorted speech segments between higher-SNR segments that stressed syllables represent. Initial results of the model suggest that a surprising degree of articulatory information is transmitted across periods during which acoustic information has been suppressed.

15. SUBJECT TERMS
Robust speech intelligibility
Computational model of speech

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Deborah Palmer
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	10	(925) 372-2343

Final Report

Project FA9550-06-1-0137 "Prosodic Stress, Information, and Intelligibility of Speech in Noise"

1. General

Speech is a process the level of which is slowly fluctuating in time (Plomp, 1983) with the result that, from the listener's perspective, some segments of running speech will appear more salient than others. This salience is modulated by prosody that produces stressed and unstressed syllables, with stress accents representing temporary increases in the level of segments that most often coincide with syllables. It has been shown that stressed syllables carry significantly more information than unstressed ones (Greenberg, Carvey, Hitchcock, and Chang, 2002). It would be therefore expected that the salient segments, when speech is presented in an interfering background (automobile or aircraft noise, cafeteria noise, crowd noise, etc.) which often also possesses a fluctuating energy, will thus have a high likelihood to be associated with an increased signal-to-noise ratio (SNR). In turn, such segments of higher-SNR will be more intelligible than the surrounding lower-SNR segments. The first objective of this project was to investigate the extent to which SNR of higher-SNR stressed syllables could account for a disproportionate amount of speech intelligibility, in an effort to broaden the scope of a first investigation showing such an effect (Divenyi, 2005).

As the results showed that unstressed speech segments gained in understanding when stressed segments had an increased SNR, the question arose what possible mechanisms could be responsible for creating the effect resembling a decrease in SNR of the unstressed syllables would have done. The phenomenon was not unlike the one of phonemic restoration (Warren, 1970) in which a speech segment substituted for by noise is not only understood in most cases but also appears, for some listeners, as if it had been present under the noise. Such a perceptual continuity suggests that some low-rate functions underlying speech continue their progression along a trajectory, as if moved by inertia, across periods during which the speech signal is either severely masked or altogether absent. This idea prompted a two-year computational work (still underway) on a model that decomposes speech into much lower-rate articulatory gesture functions and test the model to see what portion of information conveyed by these functions remains available when the speech input is halted, or replaced by some non-informative signal, for a certain period of time. This work was mainly performed by computer scientist Adam Lammert who was an assistant at the Speech and Hearing Research Lab between two graduate school programs (he is currently Annenberg Predoctoral Research Fellow at the Department of Computer Science at the University of Southern California). This research involved superimposing acoustic and articulatory-gestural representations of speech (the latter achieved by implementing the Haskins-Yale "TaDA" Task Dynamic articulatory synthesizer) and creating a codebook from a moderately large corpus. The codebook was utilized by a dynamic programming algorithm for finding the most plausible path of a given gesture predicted to underlie a test item. The method and preliminary results were presented in 2008 at an international symposium in Australia. Tests of the model were performed with a corpus of spondee words the middle portion of which were removed and replaced by silence, noise, or a voice-like periodic signal. Other tests examined differences between articulatory gesture predictions by this model and by those of other investigators.

The model's success in predicting low-rate changes in functions that underlie speech (because they can be used to synthesize speech) are in line with the recent trend of relating speech perception and production to low-frequency oscillations in the brain, as debated during a workshop on this relationship, sponsored by the AFOSR and co-directed by the PI. One unexpected outcome of the model was that it allowed analysis of listener confusions along the time axis, rather than just as a tabulated percentage of phonemes, phonetic features, or words correctly identified. By re-synthesizing the stimulus word and synthesizing the subject's response, using the TaDA synthesizer, and dynamically time-warping the response to make it comparable with the stimulus, it was possible to follow the course of the correspondence of all stimulus-response gesture pairs and to see the points of time at which the two diverged. In other words, incorrect responses could be analyzed in terms of a distance measure between a gesture in the stimulus and the same gesture in the response. Interestingly, the distances were not excessive even during the 2-300 ms intervals where the speech information was missing, suggesting that the above mentioned continuity mechanism may be active when listening to degraded or distorted speech. Furthermore, it was recognized that the distance measure could be used to propose an operational definition to the point of intersection of bottom-up auditory-sensory and top-down higher order-linguistic processes that interact to generate an erroneous response of the listener: conducting an analysis of the size of the phonetically resembling neighborhood of word alternatives could point to a stimulus-response gesture distance beyond which intervention of linguistic processes should become necessary. This idea was presented at an invitational-only session of the Acoustical Society of America's and European Acoustic Association's joint meeting in 2008.

In summary, the research conducted under the AFOSR contract examined a particular issue related to speech intelligibility under adverse circumstances: mechanisms permitting recovery of missing or severely degraded speech segments. Its relevance to combat and civilian situations of speech communication is undeniable.

2. Procedures

The research was conducted at the Speech and Hearing Research Laboratory of EBIRE. This laboratory houses two sound-attenuated chambers enclosed in a room in the middle of a building. One subject is seated in each of the chambers and listens to sounds through earphones (Sennheiser 450) and give his/her response by pressing a key on specially constructed boxes or typing it on a computer keyboard. The sounds w generated digitally by PC-type, Macintosh, or server-type computers connected in a high-speed local network running throughout the lab. One dedicated computer assigned to each of the sound rooms runs the experiment program and fetches digital audio files from dedicated disk storage systems. The digital signals are generated mostly off-line at a 44.1 kHz sampling rate converted to analog by two Echo Gina D/A systems, low-pass filtered, attenuated, and amplified using Tucker and Davis equipment, before reaching the listener's earphones.

In all experiments conducted under this project, subjects were normal-hearing young (19-30-years old) native American English speaking individuals recruited from local colleges and were paid hourly wages for their participation. They underwent the consent procedure approved by a local accredited Institutional Review Board after which their hearing was screened by the laboratory's research audiologist. They were tested for one-to-two hours at each session, in two-to-three sessions per week over a one-to-four month period, depending on the need and their availability. Overall presentation level was in the 60 to 80 dB SPL range, making it sure that peak presentation level never exceeded 90 dB SPL. All data were analyzed by locally written MATLAB routines.

3. Results

3.1 Effect of prosodic stress on speech intelligibility in noise

In these experiments the Harvard-IEEE corpus of 72 lists of ten sentences was used as the stimulus. Noise was white noise having the amplitude spectrum of the averaged spectrum of 40 sentences in the corpus. The objective of the experiment was to assess intelligibility of keywords in the sentences when, in addition to the overall SNR, the SNR of stressed syllables alone was parametrically varied. For this manipulation to be possible, a fairly sophisticated MATLAB program was written that recognized stressed syllables and put temporal markings preceding the onset and following after the coda of these syllables. The program that ran the experiments took these markings and used them to amplitude-modulate the masking noise during the periods between them by the envelope of the speech signal multiplied by a fixed coefficient set by the experimenter. (Pilot listening indicated that markings 30 ms prior to and 30 ms after the end of the syllables produced the most effective masking or unmasking, depending on the multiplier.)

Ten subjects participated in the experiment. After undergoing a two-hour training session, in each condition they were presented 60 sentences they had to respond to by typing what they had heard. Each subject heard each sentence only once, in order to avoid memory effects. A computer program parsed their responses, corrected obvious typos, and recorded them. To make it sure that the program's output was correct, the experimenter verified it before another program aligned the stimulus and response sentences and counted the keywords (five in each sentence) correctly identified. (Verb tense, article, or singular/plural noun errors were ignored.)

Pilot experiments suggested that two overall SNRs, 0 and -2 dB, were likely to yield results for young listeners in the 20 to 80 percent correct range, i.e., in the steep portion of the psychometric function. Three conditions modifying the SNR of the stressed syllables were used: one leaving it unmodulated, and two modulated conditions making the SNR of only the stressed syllables 0 dB or -2 dB. Results averaged for the ten subjects are displayed in Table 1. They show that making the stressed syllables' SNR less favorable depressed intelligibility of the sentences. Because the results were on the high side, modulation making the stress syllables more audible was not attempted with this subject group and will be postponed to a further phase of the study.

Table 1
Average percent correct keyword intelligibility (and standard errors of the mean) of ten normal-hearing native American English speaking young subjects

Stressed Syllables		Baseline SNR	
		0 dB	-2 dB
	Unmodulated	83.32 (3.44)	69.52 (5.43)
Modulated w/SNR	0 dB	83.25 (3.87)	79.48 (3.58)
	-2 dB	79.48 (3.58)	71.65 (4.63)

3.2 Decomposition of speech into articulatory gesture functions – a computational study

Results of 3.1 show that information in speech segments of unfavorable SNR can be recovered by the listener on the basis of information of surrounding higher-SNR segments. This result suggests that a continuity of some derivate of speech that changes at a slow rate may exist and may transmit information even when the acoustic signal itself is unable to, due to excessive energetic masking. One obvious candidate for this derivate is top-down linguistic information, while a different, more parsimonious alternative is represented by articulatory information that the auditory-perceptual system may be able to extract from the acoustic signal of speech. (Pointing to the plausibility of perception of articulatory functions are two facts. First, fMRI and MEG studies demonstrate activity in cortical speech motor areas when listening to speech. Second, even congenitally blind infants learn to imitate speech sounds they hear from their mother and other family members.) It was therefore attempted to develop a model of speech by decomposing it into articulatory gesture functions. Because such an attempt represents an inversion of the articulatory-to-acoustic transform which, as it has been demonstrated by many, is not possible to do directly, the model chosen started with an articulatory synthesizer, the Haskins Laboratories' Task Dynamic synthesizer (TaDA), i.e., a forward transform. From hundreds of words and a large number of sentences a codebook was built containing articulatory representation of the exemplars in form of eight continuous gesture functions (reflecting the time-varying state of eight principal articulators) aligned with their acoustic representation in form of 13 mel-frequency cepstral coefficients (MFCC) of 30-ms windowed segments. After building this codebook, it was tested by trying to determine the eight gesture functions underlying a test utterance absent from the codebook. This was accomplished by a dynamic programming algorithm that computed the probability that a the MFCC pattern of the test item probed at a given time point is associated with a given gesture having a given magnitude, based on the joint probabilities of MFCC patterns and gesture magnitudes in the codebook. The algorithm traced the predicted gesture trajectory over the duration of the test item by finding the highest-likelihood gesture magnitudes for a trajectory the smoothness of which could be broken at a minimal loss. The mathematical motivation and logic of the model was described in a paper presented at a special symposium, enclosed in the appendix of the present report. An example of the algorithm finding the trajectory of a selected gesture of a test word is shown in Figure 1.

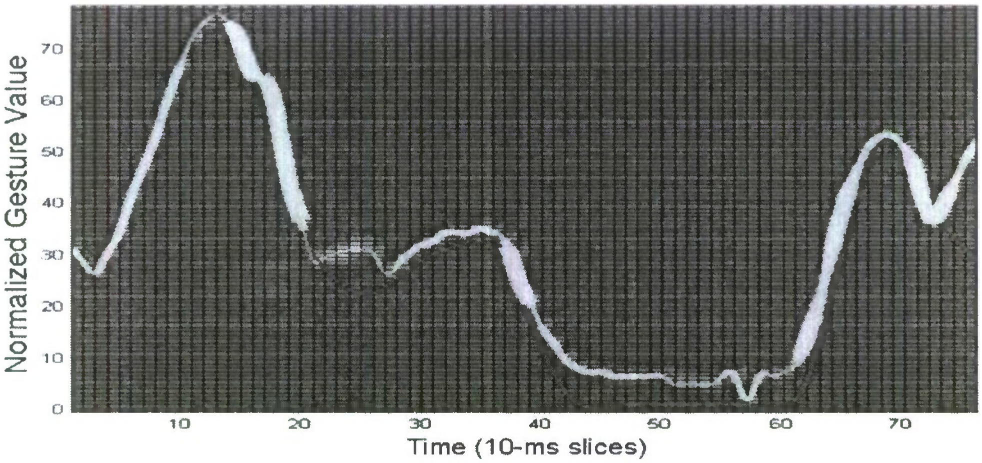

Figure 1: Trajectory of the gesture "Tongue Tip Constriction Location" associated with the synthesized word "Greymouse." The background cells represent the codebook-based probability (high: yellow – low: green) that the gesture has a certain value at a particular time slice, given the acoustic (MFCC) pattern at the same time. The blue line represents the predicted trajectory (thickness represents uncertainty) and the red line is the actual gesture trajectory generated by the synthesizer.

Clearly, as it stands, the model relies on synthesized, rather than natural, speech – a shortcoming that subsequent work plans to deal with. It is also likely that discrepancies between predicted and actual gestures will be reduced by increasing the size of the training set, i.e., by making the codebook much larger and more comprehensive. Nevertheless, despite its tentativeness, the model represents an attempt of acoustic-articulatory transform comparable in its performance to that of other investigators (e.g., Richards, Mason, Hunt, and Bridle, 1995).

3.3 Intelligibility of disyllabic words with word-middle information suppressed

Decomposition of speech into articulatory gestures may also offer a way of tracking the time course of speech recognition by human listeners, i.e., a measure of performance on the time line of the utterance to be understood. Such a timeline-bound measure of correctness could complement the discrete item- (i.e., phoneme- or phonetic feature-) bound confusion matrices. This idea was tested by the analysis of perceptual confusions in the results of a simple word recognition experiment. The stimuli were disyllabic spondee words (both real disyllabic words and words created by concatenating two unrelated real monosyllabic words) with both syllabic containing long vowels. The middle portion of the spondees, stretching from the center of the first to the center of the second syllable's nucleus, was eliminated and replaced either by silence or a filler (flat noise, noise modulated by the envelope of the cut-out speech, or a low-pass harmonic complex having the fundamental frequency contour of the cut-out speech). Fourteen subjects participated. They were asked to listen to the word and type what they thought the word was, making it clear to them that both halves were real English words.

Confusion matrices tabulated from the results were as expected: Vowels were almost never missed (since a portion of both vowels was left intact) and word-initial and word-final consonants were also most often correct; word-central (i.e., cut-out) consonants were only about 20 percent correct and even distinctive features (voicing, manner, place) were unable to reach the 50 percent mark. Interestingly, however, the eight gestures underlying the stimulus and the response words (both synthesized and time-aligned) exhibited a divergence much lower than the divergence (i.e., percent incorrect) of the word-middle phonemes. The composite divergence (Euclidean distance) of each of eight gestures in 1483 stimulus-response word pairs is shown in Figure 2.

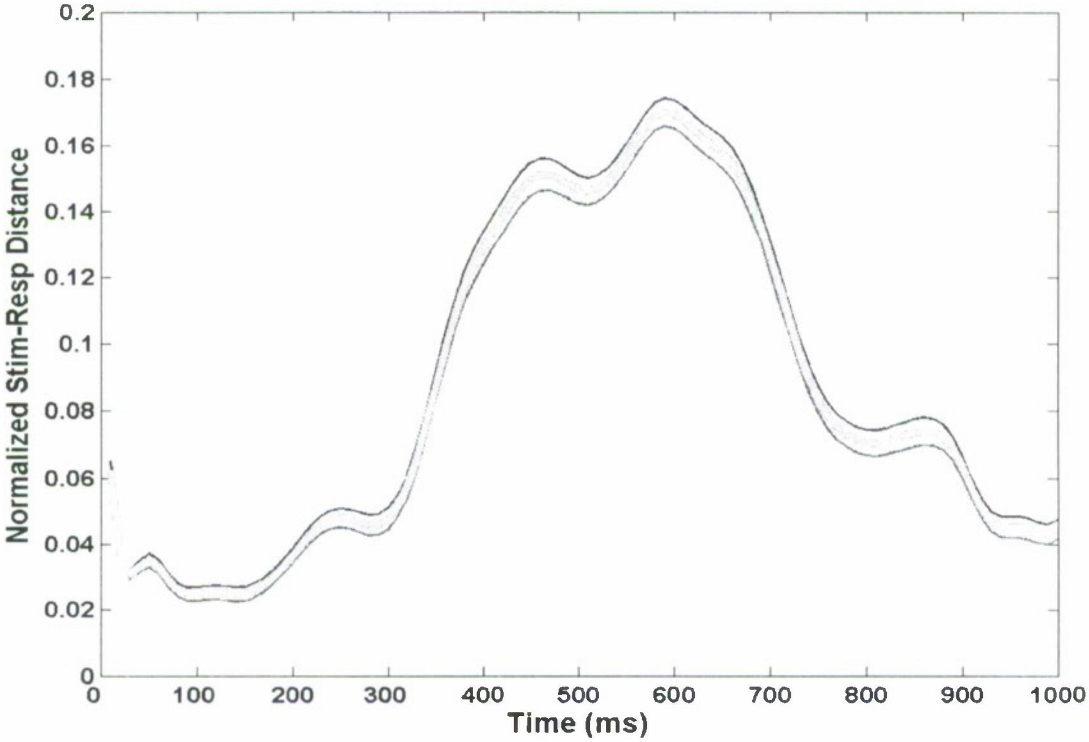

Figure 2: Normalized stimulus-response spondee word Euclidean distance averaged across eight gestures and 1483 words presented to 14 listeners. The middle of the spondees (from the center of the first to the center of the second vowel) was replaced by silence. The green line represents the mean and the two blue lines the errors (i.e., the standard error of the mean. Note that the maximum distance, in the middle of the spondee that was actually silent, is only about 18 percent, as opposed to the 80 percent rate of incorrect consonants seen in the confusion matrix for the same data.

An interesting possibility offered by such a time-bound analysis of perceptual errors is that one could, arbitrarily or guided by word statistics, fix a distance threshold beyond which one would need to invoke intervention of linguistic knowledge to make the listener choose a response word, while below that threshold a possibly more sensory-based process that tracks slowly changing functions of the speech signal could drive the decision process. One does not even need to tie such a sensory process to familiarity with articulation (although that could be the assumption easiest to defend) as long as the low rate of change in some underlying function is maintained. On a computational level, decomposition of moving images into basis functions has been shown that the basis functions can lead to a re-synthesis of the same moving images (Olshausen, 2003). Thus, the auditory equivalent of this process is ready to be investigated.

3.4 Plans for future research

The three years of the present project opened interesting alleys but the length of the grant was insufficient to go more in detail into any of the questions investigated. One reason is that the infrastructure needed to address the problems was necessary to build: programs to index and otherwise deal with large speech databases, to generate stimuli for listening experiments, to run the experiments, to analyze the data, to take the first steps toward a novel model, etc. This infrastructure in place allows not only to perform the experiments that had to be halted at the end of the grant period but also to go several steps beyond the aims of the current project. The most salient points of plans for further research are:

❖ Extend the research on prosodic effects on intelligibility to other, more complete speech corpora, like the TIMIT corpus which includes different speakers and dialects and which is more everyday English than the IEEE corpus used so far.

❖ Explore the usefulness of the gesture model for the analysis and intelligibility of speech in different types of real life noise.

❖ Explore implications of a model of speech represented by low-rate basis functions, especially as they pertain to increasing the robustness of speech understanding, speech source separation, and recognition, both in military and civilian environments.

Executive summary

Cumulative Personnel List

Pierre Divenyi, Ph.D., Principal Investigator	March 2006 – September 2008
Kara Haupt, M.S., Research Audiologist	March 2006 – September 2008
JC Saunder, B.S., Lab Assistant	March 2006 – August 2006
Joanne Hanrahan, B.A., Lab Manager	March 2006 – July 2006
Adam Lammert, M.S., Computing Specialist	March 2006 – August 2008
Casey Knifsend, B.S., Lab Manager	June 2006 – July 2008
Nicholas Livingston, M.S., Computing Specialist	July 2008 – September 2008
Nabeel Rahman, B.S., Lab Manager	July 2008 – September 2008

Publications

Articles and book chapters

Divenyi, P. (2009). Perception of complete and incomplete formant transitions in vowels. *Journal of the Acoustical Society of America*, (revision submitted).

Divenyi, P., and Lammert, A. (2007). The time course of listening bands. In B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Eds.), *Hearing - From sensory processing to perception* (pp. 175-182). Berlin, Heidelberg (Germany): Springer Verlag.

Conference presentations

Divenyi, P. (2006). *Effects of Stress Accent and Phonetic Features on the Intelligibility of Speech in Noise.* Paper presented at the Association for Research in Otolaryngology, 2006 Midwinter Meeting, Baltimore, MD.

Divenyi, P. (2007). *Décomposition temporelle des signaux de parole utilisant des fonctions de gestes.* Paper presented at the Ecole Recherche Multimodale d'Information Techniques & Sciences (ERMITES), Giens, France.

Divenyi, P., Lammert, A., and Shinn-Cunningham, B. (2008). *Perception of gestural information in words with deleted sections.* Paper presented at the 2008 Midwinter Meeting of the Association for Research in Otolaryngology, Phoenix AZ.

Divenyi, P. (2008). Localization-based segregation of acoustic sources: Advantages and limitations *J Acoust Soc Am, 123*(5), 3417 (A).

Divenyi, P., and Lammert, A. (2008). Do we perceive articulatory gestures when we listen to speech? *J Acoust Soc Am, 123*(5), 3179 (A).

Invitational conferences attended and invitational lectures given by PI

Workshop on New Ideas in Hearing, Ecole Normale Supérieure, Paris, France, May 12-13, 2006

Lecture at Oticon Research Center, Erisksholm, Denmark, August 24, 2006

Computational Auditory Scene Analysis Seminar, Danish Technical University, Lyngby, Denmark, May 24, 2007

Summer School "Ecole Recherche Multimodale d'Information Techniques & Sciences", Giens, France, September 4-6, 2007

Workshop on Temporal Dynamics in Speech and Hearing, Antwerp, Belgium, August 26, 2007

Talk at Boston University Center for Hearing Research Seminar series, December 6 2007

Talk at International Computer Science Institute, University of California, Berkeley, June 24, 2008

References cited

Divenyi, P. (2005). *Humans glimpse, too, not only machines (hommage à Martin Cooke).* Paper presented at the Forum Acusticum 2005, Budapest, Hungary.

Greenberg, S., Carvey, H. M., Hitchcock, L., and Chang, S. (2002). *Beyond the phoneme A juncture-accent model for spoken language.* Paper presented at the Proceedings of the Second International Conference on Human Language Technology Research.

Olshausen, B. (2003). *Learning sparse, overcomplete representations of time-varying natural images.* Paper presented at the IEEE International Conference on Image Processing, Barcelona, Spain.

Plomp, R. (1983). Perception of speech as a modulated signal. In M. P. R. van der Broecke, and A. Cohen (Eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences* (pp. 29-40). Dordrecht: Foris.

Richards, H. B., Mason, J. S., Hunt, M. J., and Bridle, J. S. (1995). *Deriving articulatory representations of speech.*

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science, 167,* 393-395.

Appendix

Project FA9550-06-1-0137 "Prosodic Stress, Information, and Intelligibility of Speech in Noise"

Data-driven articulatory inversion incorporating articulator priors

Adam Lammert[1], *Daniel P.W. Ellis*[2], *Pierre Divenyi*[1*]

[1] EBIRE, Martinez, CA 94553

[2] Columbia University,NewYork,NY 10027

{alammert,pdivenyi}@ebire.org, dpwe@ee.columbia.edu

Abstract

Recovering the motions of speech articulators from the acoustic speech signal has a long history, starting from the observation that a simple concatenated tube model is a reasonable model for the origin of formant resonances. In this work, we take a different approach making minimal assumptions about the interdependence of acoustics and articulators by estimating the full joint distribution of the two spaces based on a corpus of paired data, derived from an articulatory synthesizer. This approach allows us to estimate posterior distributions of articulator state as well as finding the maximum-likelihood trajectories. We present examples comparing this approach to a related, earlier approach that did not incorporate prior distributions over articulator space, and demonstrate the advantages of learning the models from realistic utterances. We also indicate benefits available from jointly estimating particular pairs of articulators that have high mutual dependence. Index Terms: articulatory inversion, speech acoustics

1. Introduction

Over the years, there have been many attempts to recover the motions of speech articulators from the acoustic speech signal. This type of recovery is an instance of the often challenging inverse problems. As such, the task is to infer model parameters from observed data. Early attempts emphasized analytical approaches, most of which sought a unique solution using acoustic models of the vocal tract [1, 2]. However, these approaches were quickly stunted, primarily by the fact that this inverse problem, like many others, suffers from non-uniqueness, and is therefore ill-posed [3]. Specifically, it is possihle for a variety of articulator configurations to produce the same acoustic result an obvious challenge for inversion. When given an acoustic signature, there will almost certainly be some ambiguity as to which of several articulator configurations produced it. This non-uniqueness has been demonstrated by many researchers over the last few decades, both from computational approaches [4,5] and experimental data[6,7].

Despite the challenges, many researchers have nonetheless forged ahead. Indeed, there is a lot of motivation to provide satisfactory estimates of speech articulators from acoustics; such results would be useful in a variety of applications [8]. Most acknowledge the trouble presented by non-uniqueness, and attempt to overcome those difficulties by using probabilistic approaches. Specifically, a variety of machine learning techniques has been applied in hopes of capturing a mapping, hoth forward and inverse, between articulator motions and acoustic observations. This kind of research has made extensive use of articulatory codebooks [9, 10, 11], as well as neural networks [12, 13, 14, 15, 16, 17]. Other studies have reported the application of dynamical models [18, 19] and stochastic techniques [8, 20, 21], including Hidden Markov Models [22]. All have seen moderate success, usually with some variability depending on the type of speech sounds being inverted. For instance, vowels and other quasi-stationary portions of speech tend to produce more successful estimations than do transitional sounds, exemplified by stop consonants.

Worth noting, additionally, are some approaches inspired by human language development [23, 24, 25]. Most of these do not explicitly address the inverse problem, but they nonetheless contribute some useful ideas. Indeed, we all know that the inverse problem is solved regularly by human beings in the first years of life, as they learn to speak. Children commonly learn to imitate speech sounds hy processing acoustic inputs, manipulating their own articulatory parameters and by drawing a mapping between the two domains. Thus, it must be possible; how can a computer be instructed to do it?

Presently, our approach uses an articulatory codebook. Codebooks are built from large sets of articulatory data with matching acoustic data, representing some sort of relation hetween the articulatory and acoustic domains. One of the first applications of codebooks to this problem used data gathered from Electromagnetic Articulography (EMA), which had been vector-quantized. A codebook assembled from this data was then used as a simple one-to-one lookup table [9]. However, using the codebook this way ignores the many-to-one mapping which makes articulatory inversion non-unique. A somewhat more successful attempt assumed a many-to-one mapping, by compiling a codebook that exhaustively covered the feasible articulatory space, using a speech synthesizer to construct the acoustic correlates of each configuration. Consequently, multiple articulator configurations could be associated with the same or very similar acoustic realizations. After compiling these options, they used dynamic programming to construct the most likely path through the field of possibilities [11]. This type of approach produced moderately successful results. A similar method was later tried [10], using data also gathered from EMA. However, to show any appreciable improvement over [11], they were required to augment their distance metrics a priori with phonemic information about the utterances.

We propose a new codebook method, which extends some previousideasand constitutesageneralizationofprior attempts. We assume that a very complex mapping may exist, and that no a priori knowledge about the utterance is available. Section 2 describes the theoretical foundation of our approach, and in

[*] This work was supported by NSF (Grant IIS-0535168), by AFOSR (Grant FA9550-06-1-0137), and the VABiomedical Laboratory Research. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, AFOSR, or the VA.

section 3 we describe our experimental implementation. We discuss the implications of these initial results in section 4.

2. Approach

In this prior work, the most common assumption is of a deterministic acoustic system – which is to say if the articulator positions $\mathbf{A}$ are known, the acoustic observations $\mathbf{O}$ can be directly and unambiguously determined through some (nonlinear) function, $\mathbf{O} = f(\mathbf{A})$. Observations may not completely determine the articulators i.e. there may be several values of $\mathbf{A}$ that result in the same $\mathbf{O}$, but the forward acoustics are unambiguous i.e. there is only one $\mathbf{O}$ for a given $\mathbf{A}$. If, however, we have a system where the articulator state $\mathbf{A}$ is incomplete, then we may have a doubly-ambiguous situation, where a single $\mathbf{A}$ can result in multiple values for $\mathbf{O}$, as well as vice-versa. In this case, a more appropriate way to describe the relationship between articulator state and acoustic observations is as a joint probabilistic distribution $p(\mathbf{A}, \mathbf{O})$, which simply describes the absolute likelihood of any combination of articulator state and acoustic observation. Such a probabilistic description could also incorporate a number of other aspects of the problem, including measurement uncertainty, unmodeled variability in the system, and the *a priori* probabilities of particular acoustics and particular articulatory configurations (independent of each other). Given a suitable approximation to $p(\mathbf{A}, \mathbf{O})$, the inverse acoustics problem of inferring articulators $\mathbf{A}$ from acoustic observations $\mathbf{O}$ amounts to calculating the posterior distribution of the articulators given the observations i.e.:

$$\begin{aligned} p(\mathbf{A}|\mathbf{O}) &= p(\mathbf{A}, \mathbf{O})/p(\mathbf{O}) \\ &= \frac{p(\mathbf{A}, \mathbf{O})}{\int_{\mathbf{A}} p(\mathbf{A}, \mathbf{O}) d\mathbf{A}} \end{aligned} \tag{1}$$

The well-known articulatory ambiguity for given acoustics would emerge as a broad and/or multimodal posterior distribution for the articulators. A time-local model of acoustics and articulators could then be disambiguated by continuity considerations e.g. using dynamic programming to find the best complete path through a sequence of articulator posterior distributions.

This is the approach we take. $p(\mathbf{A}, \mathbf{O})$ should encompass the full range of articulatory and acoustic states anticipated in natural speech, in the appropriate proportions (i.e. with the greatest likelihood for the most common speech sounds and their most common articulatory counterparts). If we had an unlimited database of real speech, along with the true underlying articulator positions, we could simply sample from this database at random, until we had enough points to provide an adequate sampling density in the joint feature space, then perform some kind of local density estimation and smoothing to approximate $p(\mathbf{A}, \mathbf{O})$ – for instance by 'blurring' each distinct articulator-acoustics pair to account for a small range of local values in both domains. This amounts to Parzen estimation [26].

In the absence of a large, representative database of actual articulator and acoustic pairs, we used a forward articulatory speech synthesis model to construct a range of more or less natural and phonetically-balanced sentences (drawn from the Harvard IEEE set). To the extent that this training data includes all the main speech sounds (and articulator configurations), this database should be adequate to model the joint distribution of articulators and acoustics, at least for the specific 'vocal system' being modeled by the synthesizer. Thus, if we define a set of distance vectors between a given point in the joint articulatory-acoustic space $(\mathbf{A}, \mathbf{O})$ and each of our training examples,

$$\Delta_i = \begin{bmatrix} \mathbf{A} - \mathbf{A}_i \\ \mathbf{O} - \mathbf{O}_i \end{bmatrix} \tag{2}$$

where $\{\mathbf{A}_i, \mathbf{O}_i\}$ are our N training patterns, we can approximate our joint distribution,

$$p(\mathbf{A}, \mathbf{O}) = \frac{|\mathbf{W}|}{N} \sum_{i=1}^{N} K\left(\Delta_i^T \mathbf{W} \Delta_i\right) \tag{3}$$

where $\mathbf{W}$ is a positive-definite weighting matrix that defines the 'width' of the Parzen smoothing windows in the articulator and acoustic feature spaces. $K(\cdot)$ is the Parzen window itself, for instance a unit-variance normalized Gaussian, $K(\alpha) = 1/\sqrt{2\pi} \exp\{-\alpha^2/2\}$.

The choice of $\mathbf{W}$ depends on the sparsity of the sample density in each dimension as well as assumptions about the smoothness of the joint distribution along those dimensions. Taking $\mathbf{W}$ as diagonal, a small entry in a particular dimension corresponds to a wide window in that dimension, allowing for density to be interpolated between relatively broadly-spaced samples, but at the same time smoothing out any variation in more densely-sampled regions of the space that occurs at a finer scale. One adaptive approach to this is to vary the effective window width in proportion to the local density – for instance, by finding the k nearest neighbors to a given point, then setting the window width at that point as some fixed factor times the average distance to these neighbors, and performing this separately for each dimension.

In practice, then, we can calculate posterior distributions for articulatory parameters (either as a group, or as subsets in which case unused dimensions are ignored) by taking the acoustic observations $\mathbf{O}$, then retrieving all the training patterns $\{\mathbf{A}_i, \mathbf{O}_i\}$ within the radius of support of the Parzen window K over the acoustic dimensions. Then, for a each value in a grid defined over the possible values for the articulators $\mathbf{A}$, the joint probability $p(\mathbf{A}, \mathbf{O})$ of the actual observation and the hypothesized articulator value is calculated via eqn. 3. Normalizing by the sum over all articulator values gives the posterior probability (according to eqn. 1) — although since the subsequent dynamic programming search is obliged to choose exactly one articulator value for each time step, a common scaling of all likelihood scores at a particular time will not change the optimal choice, and thus the normalization is not required in practice.

In the case of independent estimation of a single articulatory parameter, the result of this is that a sequence of acoustic observation vectors results in a table of joint probabilities with each row corresponding to one of the quantized possible articulator values, and each column corresponding to one time step. We can regard the columns as sets of scaled posteriors for the articulatory parameters, and use dynamic programming to find the most likely sequence of articulatory values by simultaneously applying a continuity constraint as a transition cost that penalizes large jumps in articulator position. Specifically, we estimate the sequence of articulator values $\{\hat{\mathbf{A}}_t\}$ by using dynamic programming to find the sequence that maximizes

$$\prod_t p(\hat{\mathbf{A}}_t|\mathbf{O}_t) q(\hat{\mathbf{A}}_t|\hat{\mathbf{A}}_{t-1}) \tag{4}$$

where $q(\hat{\mathbf{A}}_1|\hat{\mathbf{A}}_0)$ is defined as 1, and

$$q(\hat{\mathbf{A}}_t|\hat{\mathbf{A}}_{t-1}) = \exp\left\{-\frac{1}{2}\left(\left|\hat{\mathbf{A}}_t - \hat{\mathbf{A}}_{t-1}\right|/\sigma\right)^2\right\} \tag{5}$$

for $t > 1$. σ is taken as the 99th percentile of articulator first-order differences seen in the training data, and is calculated separately for increasing and decreasing changes. Joint estimation of multiple articulator dimensions can be performed similarly, for as far as it is practical to track every value in the uniformly-quantized articulator space space. Practically, this has limited us to two dimensions in the current work.

A key aspect of our approach is that, by making our core model an estimate of $p(\mathbf{A}, \mathbf{O})$ (the joint density of articulators and acoustics) rather than starting from a conditional model of the probability of acoustics under different articulatory configurations $p(\mathbf{O}|\mathbf{A})$ (the approach more or less implicit in previous related work), we are modeling not only the relation between the two spaces, but also the prior likelihood of particular configurations in both spaces. This is particularly important in disambiguating articulatory states that may be roughly equally good explanations of a particular acoustic observation, but which may differ greatly in their *a priori* likelihood, on the basis of the frequency with which they occur in normal speech. In the experiments section below, we contrast three variants of this model. The first, comparison model, attempts to remove the effects of including these priors by normalizing each joint probability by the prior of associated articulation to obtain the conditional i.e.

$$p(\mathbf{O}|\mathbf{A}) = p(\mathbf{A}, \mathbf{O})/p(\mathbf{A}) \tag{6}$$

(note the difference between this and equation 1, our true objective). The normalization denominator, which is the prior probability of the articulatory configuration $p(\mathbf{A})$, is obtained simply by calculating a histogram of articulator values (over all dimensions, or some subset) over the entire training set. By reducing our joint density to a model of acoustics given articulation, we intend this variant to be equivalent to the uniform sampling of articulator space that gave the codebook of [11], which did not incorporate articulator priors.

The remaining variants use our full joint distribution, and estimate articulator values either individually and independently, or in pairs. When more than one articulator is estimated at the same time, estimation can exploit joint dependence between articulator behavior, which ought to improve performance in the cases where articulators are directly linked, but may be ambiguous when viewed independently. We will discuss our results from this perspective.

3. Experiments

Our codebook is based on synthesized speech obtained from the Task Dynamic Application (TaDA) developed at Haskins Laboratories. TaDA is a MATLAB implementation of the Task Dynamic model of speech articulator coordination [27]. As such, it uses articulator positions as basis functions from which to synthesize speech. There are eight relevant articulator positions: tongue tip constriction degree (TTCD) and location (TTCL), tongue body constriction degree (TBCD) and location (TBCL), lip aperture (LA) and protrusion (PRO), velum (VEL) and glottis (GLO). TaDA simultaneously generates input parameters for the HLSyn speech synthesis software, which synthesizes more natural-sounding speech. The speech output of HLSyn is then transformed into 13 Mel Frequency Cepstral Coefficients (MFCCs) [28], using a 10ms window size and 5ms window advance rate. The articulator positions from TaDA, matched with the MFCCs constitute the final codebook.

The codebook used in our experiments was built from a training data set composed of 40 natural and phonetically-balanced sentences, drawn from the Harvard IEEE Corpus. To input the sentences into TaDA, we used the program's capability to receive orthographic input. From within TaDA, this orthography is then converted into phonemes via a dictionary lookup procedure.

We ran several experiments in order to compare three distinct articulatory inversion methods. The first was an implementation of a previous attempt by Richards [11] and the others were two variants of our own method. Specifically, the first attempted to estimate a single articulator (labeled 1A), and the second sought to estimate two articulators simultaneously (2A). All three methods were tested on two conditions, represented by the same two testing sentences. One sentence was designed to be close to the training sentences (labeled 'Easy'), in order to mimic a larger training set. This sentence contained only words that appeared in the training set sentences. The other sentence was more novel, with only 29% of words in the training set (labeled 'Hard').

The literature on articulatory inversion provides no consensus about how to compare estimated articulator paths with actual paths. Both correlation and geometric distance measures are well represented in the diversity of studies on this topic. We chose to use simple Euclidean distance as a way to compare the actual articulator paths (as determined by TaDA) with the ones estimated by the various inversion methods. For the Richards method, there were a total of 16 distance measurements calculated (8 articulators gathered for 2 testing sentences). The same number of measurements was also calculated for the 1A method. For the 2A method there were more measurements necessary. We continued to track each articulator for each testing word, but also for each accompanying articulator. This brought the total number of measurements to 112 (8 articulators gathered for 2 testing sentences with 7 possible accompanying articulators).

3.1. Results

Tables 1 and 2 summarize the results of the different models in terms of mean squared error in the normalized articulator estimates relative to the ground truth. When looking at the results of the Richards method versus our 1A method, a substantial improvement can be seen. This can be seen even across articulators. The mean Euclidean distance for the Easy testing word improved 20.6% when using the 1A method, as opposed to the Richards method. Surprisingly, an even larger disparity was seen for the Hard testing word, where there was an improvement of 29.9% for the 1A method. Moreover, the 1A results were about the same or better for every articulator, with the exception of lip protrusion, in which 1A lowered performance by 50% on the Easy test sentence only. Conversely, the most dramatic improvement was seen with tongue tip constriction degree which improved by 67.4% on the Easy sentence.

In comparing methods 1A and 2A, an improvement can also be seen. There is some variability in the improvement which depends on the articulator pairs chosen. However, the improvement is evident in the overall case. For the Hard testing sentence, the mean across all articulators (target and accompanying) is 3.7% lower than the 1A mean. This improvement is not evident for the Easy sentence, however, which appears to be 8.3% worse upon first inspection. This is mainly due to one accompanying articulator – tongue body constriction degree – which proved to be a major hindrance to the estimation of all target articulators. If one removes these data points, then overall mean for the 2A method shows a 7.4% improvement over the 1A method. The largest improvement of articulator estimation

Table 1: Results of estimating articulator positions by different models: 'Easy' sentence. Best values in each column are in bold.

Alg.	with	GLO	LA	PRO	TBCD	TBCL	TTCD	TTCL	VEL	Mean
Rich	–	38.5	15.9	18.8	62.3	10.8	29.3	17.4	12.9	25.7
1A	–	18.8	7.3	28.2	62.3	8.2	**9.6**	15.9	13.1	20.4
2A	GLO	18.8	7.0	27.7	62.3	8.9	9.9	16.6	12.3	20.4
	LA	14.1	7.3	19.0	62.3	8.6	14.1	10.5	9.3	18.2
	PRO	18.8	8.0	28.2	62.3	9.0	12.7	14.3	12.3	20.7
	TBCD	18.8	36.3	40.9	62.3	61.6	38.4	56.6	50.8	45.7
	TBCL	12.1	6.7	21.4	62.3	8.2	14.4	16.0	10.5	19.0
	TTCD	14.4	7.2	**17.9**	62.3	9.3	**9.6**	12.2	8.8	**17.7**
	TTCL	**10.9**	6.7	19.9	62.3	9.5	14.3	15.9	**8.3**	18.5
	VEL	15.6	**6.6**	18.9	62.3	**8.1**	9.9	**9.3**	13.1	18.0

Table 2: Results of estimating articulator positions by different models: 'Hard' sentence. Best values in each column are in bold.

Alg.	with	GLO	LA	PRO	TBCD	TBCL	TTCD	TTCL	VEL	Mean
Rich	–	48.7	37.6	20.8	27.9	25.3	13.1	18.5	11.1	25.4
1A	–	31.1	28.2	19.7	12.2	16.6	11.9	13.4	9.5	17.8
2A	GLO	31.1	27.5	19.7	14.3	16.3	11.5	14.4	9.8	18.1
	LA	28.1	28.2	**19.6**	12.9	13.4	11.1	15.1	**8.5**	17.1
	PRO	31.1	26.8	19.7	16.5	18.2	11.9	12.7	8.8	18.2
	TBCD	27.5	**24.5**	19.7	12.2	14.9	10.2	**11.2**	**8.5**	**16.1**
	TBCL	28.9	26.9	19.7	14.8	16.6	**9.6**	13.4	9.3	17.4
	TTCD	26.8	27.9	19.7	**11.2**	12.6	11.9	13.6	8.7	16.6
	TTCL	**26.7**	28.1	19.7	15.5	**9.7**	12.0	13.4	**8.5**	16.7
	VEL	27.2	27.6	19.7	13.5	14.5	11.5	13.8	9.5	17.2

was seen with the pairing of glottis and tongue tip constriction location. This allowed the estimation of the glottis to be improved 42.2% over the 1A method for the Easy sentence. The largest mean improvement across target articulators was seen with the assistance of tongue tip construction degree as the accompanying articulator, which improved estimation of the articulators by 13.3% for the Easy sentence.

Thus, a complicated picture arises when considering method 2A. It was not a clear win over 1A and, even though the overall picture was positive, the specifics of the picture are mixed. Some articulators seem to be a great help as an accompaniment to other articulators, and some other articulators gain benefits from being accompanied without regard to which articulator. For instance, velum, glottis and lip protrusion seem to be aided by the majority of accompanying articulators. At the same time, tongue tip constriction degree and location and lip aperture appear to help in the estimation of most other articulators. The picture is mixed, though, as the estimation of some articulators was hindered by 2A. To tongue body constriction degree, estimation with almost every other articulator was detrimental, while glottis confused nearly every articulator it was paired with.

Figures 1 and 2 show results for the estimation of individual articulators (tongue body constriction location and lip aperture, respectively) for each of the three methods applied. For all plots, the thin red line represents the actual articulator path, while the thicker blue line represents the estimated articulator values. Behind each plot is shown the local-match scores used to determine the estimated path with the dynamic programming algorithm. These scores can also be thought of as probabilities, at a given instant in time, of the articulator taking on a particular value, given the observed acoustics at that time. The two articulators chosen were jointly estimated for the representation of the 2A case, as indicated. Thus, the 2A results are projections of the 3-dimensional estimation space which has the dimensions LA, TTCL and time. For both articulators, the 2A estimation produced superior results to either of the 1A estimations. Additionally, the 1A estimation was, for both individually, superior to the Richards estimation.

4. Discussion and Conclusions

We have presented an approach to estimating articulator configuration directly from acoustics based on a model of the joint distribution of the two, multidimensional spaces that incorporates both the link between articulators and acoustics, and the prior probabilities in both spaces, based on a paired corpus that is taken to reflect the balance of gestures in real speech. This approach can accommodate arbitrarily complex relationships, including ambiguities (multimodality) in either domain. When ambiguity occurs, valid articulation can still be inferred based both on continuity in articulator space (as enforced by transition constraints through time) and on differing priors among the alternative explanations.

Our transition modeling is somewhat deficient compared to the careful model of joint density: we use a single, global cost function to discourage large excursions in our dynamic programming best-cost paths, rather than, say, attempting to model the actual dynamics present in our training data. Indeed, it would be possible to extend the probabilistic model used at the frame level to obtain a posterior over *sequences* of articulators given sequences of observations, using the hidden Markov model (HMM). There is a problem here, however: In the conventional exposition of the HMM, local constraints are incorpo-

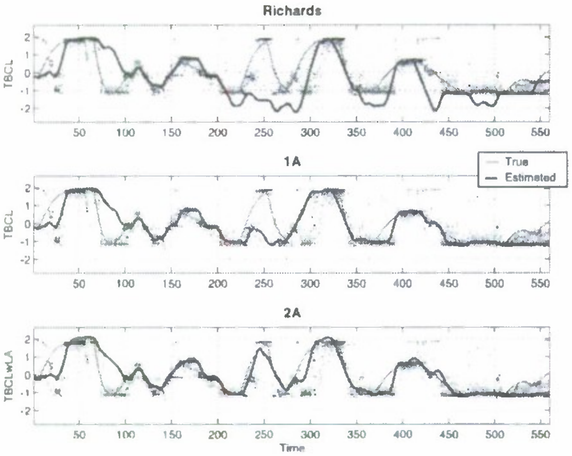

Figure 1: Example estimates for TBCL articulator from the three models. Top: Richards model (no priors). Middle: single-articulator current model. Bottom: TBCL from joint estimation of TBCL and LA. Ground truth (i.e. input to the synthesizer) is shown in each case, as is the underlying score surface input to the dynamic programming.

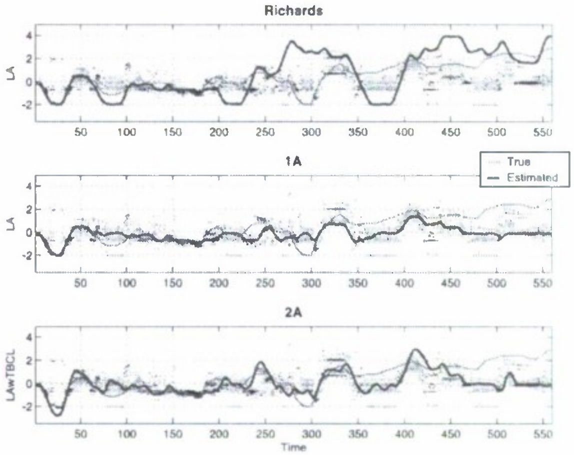

Figure 2: Example estimates for LA articulator from the three models as fig 1.

rated as the conditional distribution of observations given state, $p(\mathbf{O}|\mathbf{A})$, not the posterior probabilities of state given acoustics $p(\mathbf{A}|\mathbf{O})$ adopted as the goal here. The HMM then applies the prior of particular state sequences via the state transition costs, $p(\mathbf{A}_t|\mathbf{A}_{t-1})$ which incorporate both the likelihood of a particular transition and, implicitly, the overall likelihood of particular state configurations. One interpretation of our current approach is that we have taken state-specific variations in the transition probabilities and incorporated them in our local match scores, allowing a single, global, normalized transition cost. However, experiments to estimate and model transition behavior more accurately are an important direction for future work.

4.1. Future Work

The most significant challenges we face in the future revolve around moving toward speaker independence. Our results, no matter how much improved, are tied tightly to a specific – and in this case synthetic – speaker. The speaker that is implicitly represented by TaDA was our sole speaker for the experiments described herein. Moreover, the speech generated by TaDA is bound by a phonemic decomposition of the signal, which serves as the input to the synthesizer. Both of these factors mean that our training data, from which we build our codebook, is lacking in variation. Thus, we suspect that our codebook approach would struggle to predict the articulator motions of speech which is variable and substantially different from the TaDA speaker. That is to say, we suspect a challenge in applying our technique to additional speakers and to natural speech. There are several possibilities that could aid in overcoming this shortcoming. One idea is to average across a range of speakers. This could be accomplished by simply introducing variation into the training data in the form of new and different speakers. However, we tend to favor normalization of features into a speaker-independent space.

Of course, much rests on the assumption that we can obtain large amounts of training data in the future. From the standpoint of collecting synthetic speech, this may not pose too much a challenge. Different synthetic speakers can be created by varying the parameters of the synthesizer. Collecting speech-articulator pairing data for natural speech has been a challenge over the years. Several groups are currently working on compiling these data. Notably, the Speech Production and Articulation kNowledge Group (SPAN) is using Magnetic Resonance Imaging to capture data about the very same articulator used in TaDA [29]. This sort of endeavor is absolutely crucial to the future of solving articulator inversion.

Since to much of our methodology stands on the shoulders of the acoustic representation, it may be necessary to explore more sophisticated options for that representation in the future. Although MFCCs seem to perform well, especially in conjunction with geometric measures of similarity, they are by no means the ultimate choice. It is disappointing to note that very few representations of the speech acoustics have attempted to track the spectral changes caused by changes in the articulators, despite the fact that the articulatory functions' primary effect on the acoustics of the speech waveform consists of changing the source filter characteristics. In one of the next incarnations of our model, we will supplement or supplant our MFCC-based frame-by-frame acoustic data with information containing the magnitude and phase of changes in a finite number of frequency prominences, in the hope that such a modification will result in the acoustics being more strongly bound to the articulatory functions.

5. References

[1] P. Mermelstein and M. Schroeder, "Determination of Smoothed Cross-Sectional-Area Functions of the Vocal Tract from Formant Frequencies," *The Journal of the Acoustical Society of America*, vol. 37, p. 1186, 1965.

[2] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 417–427, 1973.

[3] E. Borowski and J. Borwein, *Dictionary of Mathematics*. Harper Collins, 1991.

[4] B. Lindblom, J. Lubker, and T. Gay, "Formant frequencies of some fixed-mandible vowels and a model of speech

motor programming by predictive simulation," *Journal of Phonetics*, vol. 7, pp. 146–161, 1979.

[5] S. Roweis, "Data Driven Production Models for Speech Processing," *Unpublished Ph. D. Thesis, California Institute of Technology, Pasadena, CA*, 1999.

[6] B. Atal, J. Chang, M. Mathews, and J. Tukey, "Inversion of articulatory-to-acoustic tranformation in the vocal tract by a computer-sorting technique," *J. Acoust. Soc. Am.*, vol. 63, no. 5, pp. 1535–1555, 1978.

[7] J. Flanagan, *Speech Analysis Synthesis and Perception*. Springer, 1972.

[8] S. Dusan and L. Deng, "Estimation of articulatory parameters from speech acoustics by kalman filtering," 1998. [Online]. Available: citeseer.ist.psu.edu/dusan98estimation.html

[9] J. Hogden, A. Lofqvist, V. Gracco, I. Zlokarnik, P. Rubin, and E. Saltzman, "Accurate recovery of articulator positions from acoustics: New conclusions based on human data," *The Journal of the Acoustical Society of America*, vol. 100, pp. 1819–1834, 1996.

[10] T. Okadome, S. Suzuki, and M. Honda, "Recovery of articulatory movements from acoustics with phonemic information," in *Proc. 5th Seminar on Speech Production*, Kloster Seeon, Germany, May 2000, pp. 229–232.

[11] H. Richards, J. Mason, M. Hunt, and J. Bridle, "Deriving articulatory representations from speech with various excitation modes," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 1233–1236, 1996.

[12] M. Huckvale and I. Howard, "Teaching a Vocal Tract Simulation to Imitate Stop Consonants," in *Ninth European Conference on Speech Communication and Technology*. ISCA, 2005.

[13] G. Papcun, J. Hochberg, T. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 92, pp. 688–700, 1992.

[14] P. Perrier, L. Ma, and Y. Payan, "Modeling the production of VCV sequences via the inversion of a biomechanical model of the tongue," in *Proc. Interspeech*, Lisbon, 2005, pp. 1041–1044.

[15] M. Rahim, W. Keijn, J. Schroeter, and C. Goodyear, "Acoustic to articulatory parameter mapping using an assembly of neural networks," *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pp. 485–488, 1991.

[16] K. Richmond, "Estimating articulatory parameters from the acoustic speech signal," Ph.D. dissertation, PhD thesis, Centre for Speech Technology Research, Edinburgh University, 2001.

[17] G. Westerman and E. Miranda, "Modelling the development of mirror neurons for auditory-motor integration," *Journal of New Music Research*, vol. 31, no. 4, pp. 367–375, 2002.

[18] K. Shirai and T. Kobayashi, "Considerations on articulatory dynamics for continuous speech recognition," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8, 1983.

[19] ——, "Estimating articulatory motion from speech wave," *Speech Communication*, vol. 5, no. 2, pp. 159–170, 1986.

[20] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," *Proc. 5th Seminar on Speech Production*, pp. 237–240, May 2000.

[21] S. King and A. Wrench, "Dynamical system modeling of articulator movement," in *Proc. International Congress of Phonetic Sciences*, San Francisco, CA, 1999.

[22] G. Ramsay and L. Deng, "Optimal filtering and smoothing for speech recognition using astochastic target model," *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 1113–1116, 1996.

[23] G. Bailly, "Learning to speak. sensori-motor control of speech movements," *Speech Communication*, vol. 22, no. 2-3, pp. 251–267, 1998. [Online]. Available: citeseer.ist.psu.edu/bailly98learning.html

[24] F. Guenther, "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production," *Psychological Review*, vol. 102, no. 3, pp. 594–621, 1995.

[25] K. Markey, "The sensorimotor foundations of phonology: a computational model of early childhood articulatory and phonetic development," Ph.D. dissertation, University of Colorado, 1994.

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition*. Wiley, New York, 2001.

[27] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "TADA: An enhanced, portable Task Dynamics model in MATLAB," *Acoustical Society of America Journal*, vol. 115, no. 5, pp. 2430–2430, 2001.

[28] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: www.ee.columbia.edu/ dpwe/resources/matlab/rastamat/

[29] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America*, vol. 115, pp. 1771–1776, 2004.